Preparing Applications for *Mira*, a 10 PFLOPS IBM Blue Gene/Q Machine

Timothy J. Williams

Argonne Leadership Computing Facility

Argonne National Laboratory

2011 Fall Creek Falls Conference

9/15/2011

- **ALCF**
- ***Intrepid***
  - Hardware
  - Software
  - How to use
- ***Mira***
- **Applications**
  - Co-Design
  - Early Science Program
  - Tools & Libraries

Argonne Leadership Computing Facility

- **Established 2006 at Argonne National Lab**
- **One of two DOE national Leadership Computing Facilities (OLCF is other)**
- **Supports mission of DOE Office of Science Advanced Scientific Computing Research (ASCR)**

DOE INCITE Program

Innovative and Novel Computational Impact on Theory and Experiment

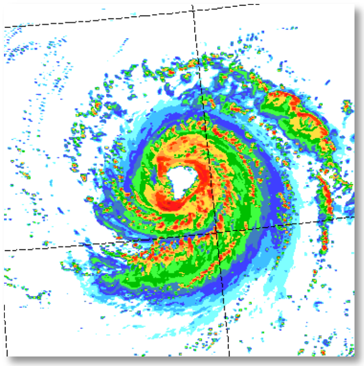
- **60% of time at Leadership Facilities**
- **Solicits large, computationally intensive research projects**
  - To enable high-impact scientific advances
  - Call for proposals yearly (closed 6/30/2011)
  - INCITE Program web site: doeleadershipcomputing.org
- **Open to all scientific researchers and organizations**
  - Scientific discipline peer review
  - Computational readiness review
- **Awards large computer time & data storage allocations**
  - Small number of projects for 1-3 years
  - Academic, national lab and industry, with DOE or other support
- **2011 INCITE at ALCF**
  - **30 projects**
  - 732**M core hours**

DOE ALCC Program

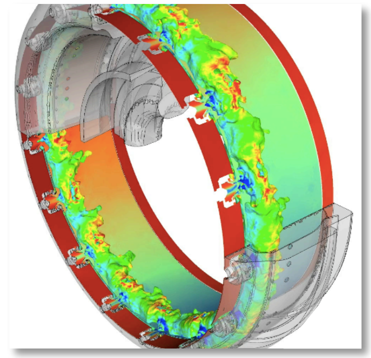
ASCR Leadership Computing Challenge

- **30% of time at LCFs**
- **Projects of special interest to DOE**
  - emphasis on high-risk, high-payoff simulations
- **Awards granted in June (review started 2/18/2011)**
  - science.energy.gov/ascr/facilities/alcc
- **2011 ALCC at ALCF**
  - 7 awards
  - 300+ million core hours

ALCF Projects Span Many Domains

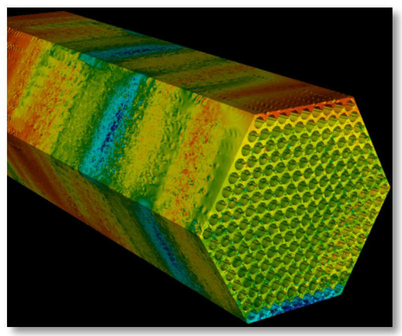


Climate
Predicting hurricane tracks to mitigate risks, hindcasting with climate model data to gauge impact of global change.

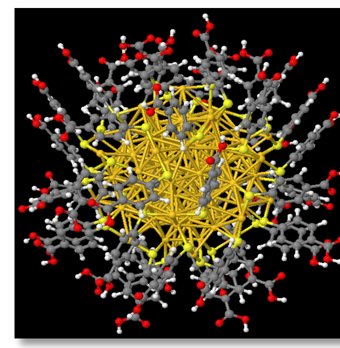


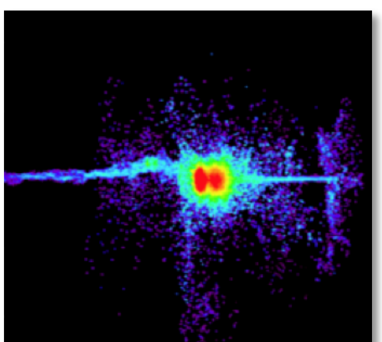
Gas Turbines
Modeling two-phase flow and combustion for the design of more efficient aircraft engines.

Nuclear Energy
High-fidelity fluid flow and heat transfer simulation of next-generation reactor designs, aiming to reduce the need for costly experimental facilities.

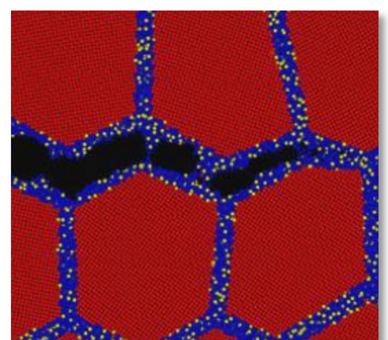


Nano Catalysts
Mapping out properties of gold nanoparticles to design catalysts for fuel cells and methane conversion.





Fusion Energy
Understanding the detailed physics of Fast Ignition inertial confinement fusion.



Materials Science
Molecular simulation of fracture dynamics in structural materials in next-generation nuclear reactors.

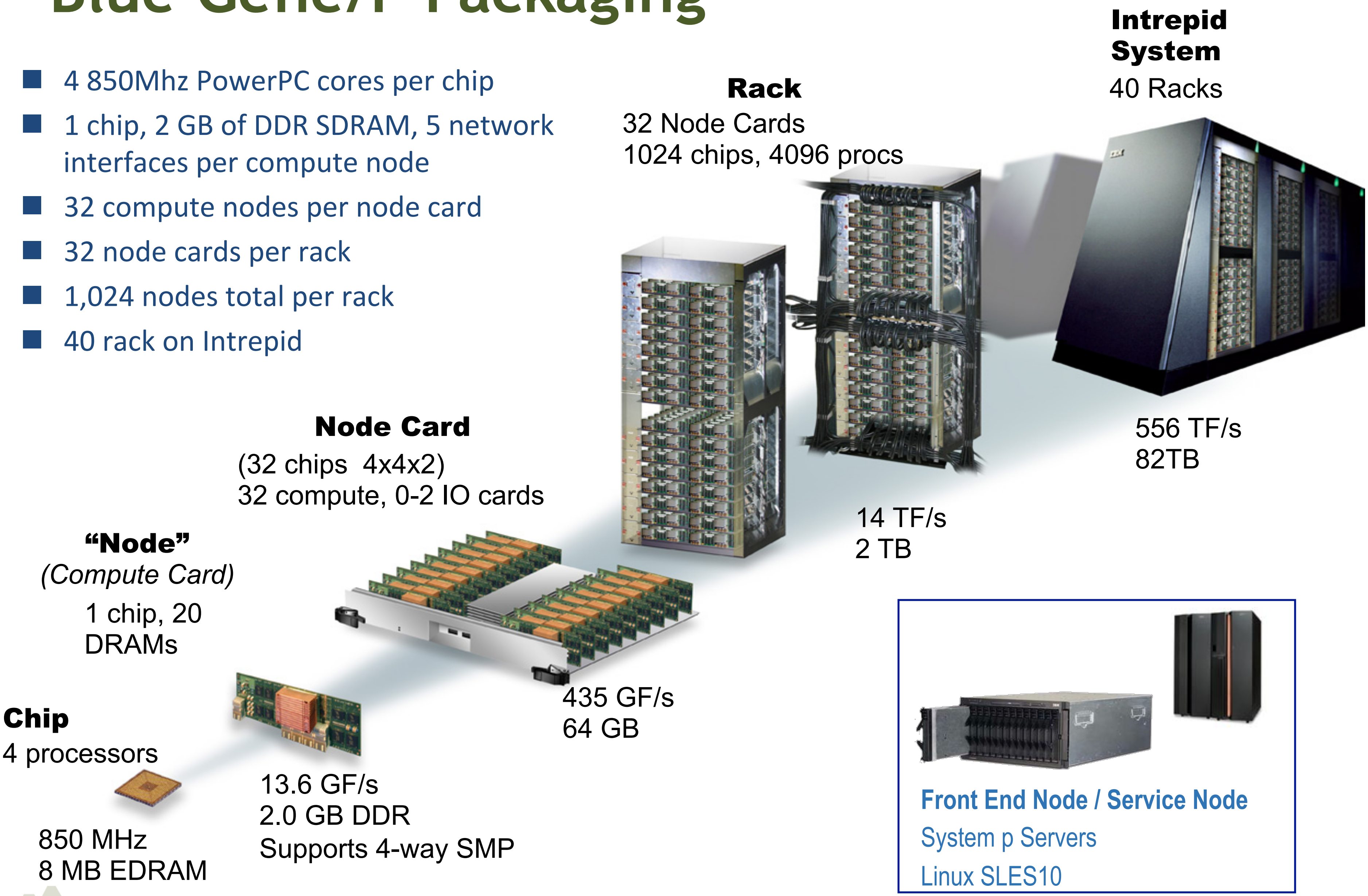
ALCF Hardware

- *Intrepid* - ALCF Blue Gene/P System:
  - 40,960 nodes / 163,840 PPC cores
  - 80 Terabytes of memory
  - Peak flop rate: 557 Teraflops
  - Linpack flop rate: 450.3
  - #15 on the Top500 list
  - #1 on Graph500 list
  - #41 on Green500 list

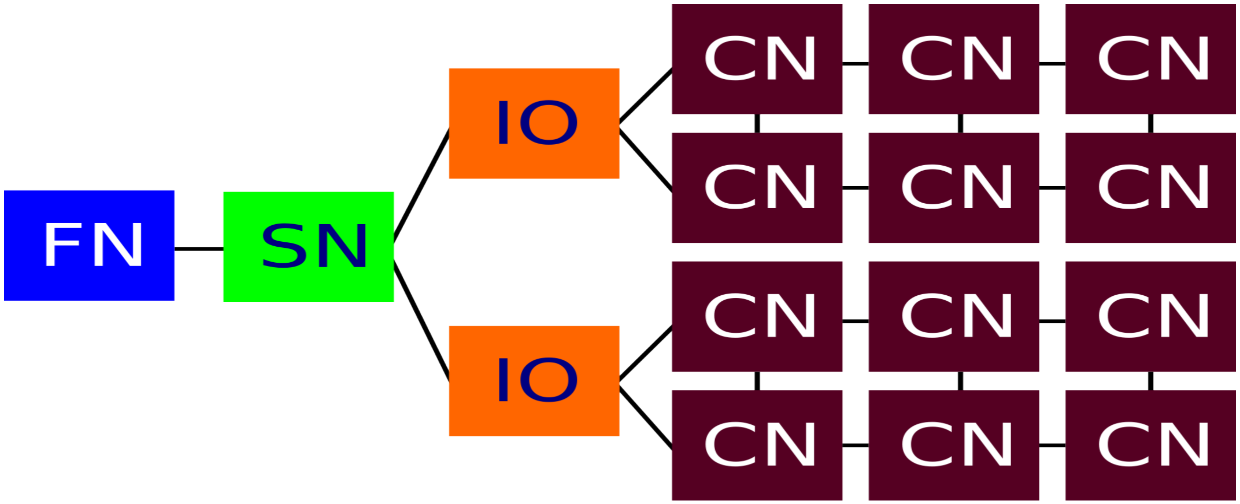
- *Eureka* - ALCF Visualization System:
  - 100 nodes / 800 2.0 GHz Xeon cores
  - 3.2 Terabytes of memory
  - 200 NVIDIA FX5600 GPUs
  - Peak flop rate: 100 Teraflops
- Storage:
  - 6+ Petabytes of disk storage with an I/O rate of 80 GB/s
  - 5+ Petabytes of archival storage (10,000 volume tape archive)

Blue Gene/P Packaging

- 4 850Mhz PowerPC cores per chip
- 1 chip, 2 GB of DDR SDRAM, 5 network interfaces per compute node
- 32 compute nodes per node card
- 32 node cards per rack
- 1,024 nodes total per rack
- 40 rack on Intrepid

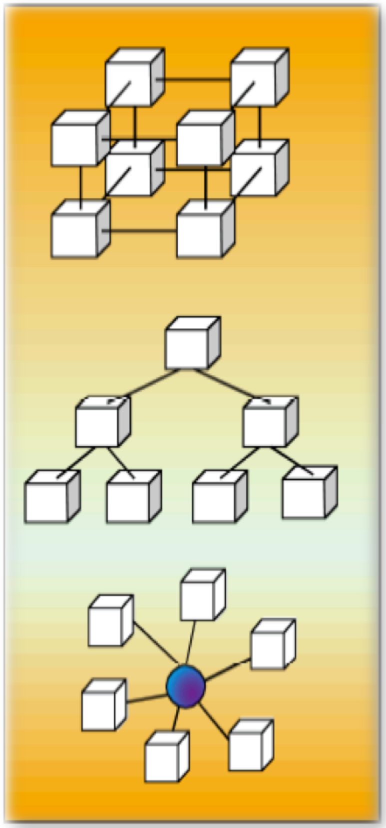


Blue Gene/P Heterogeneity



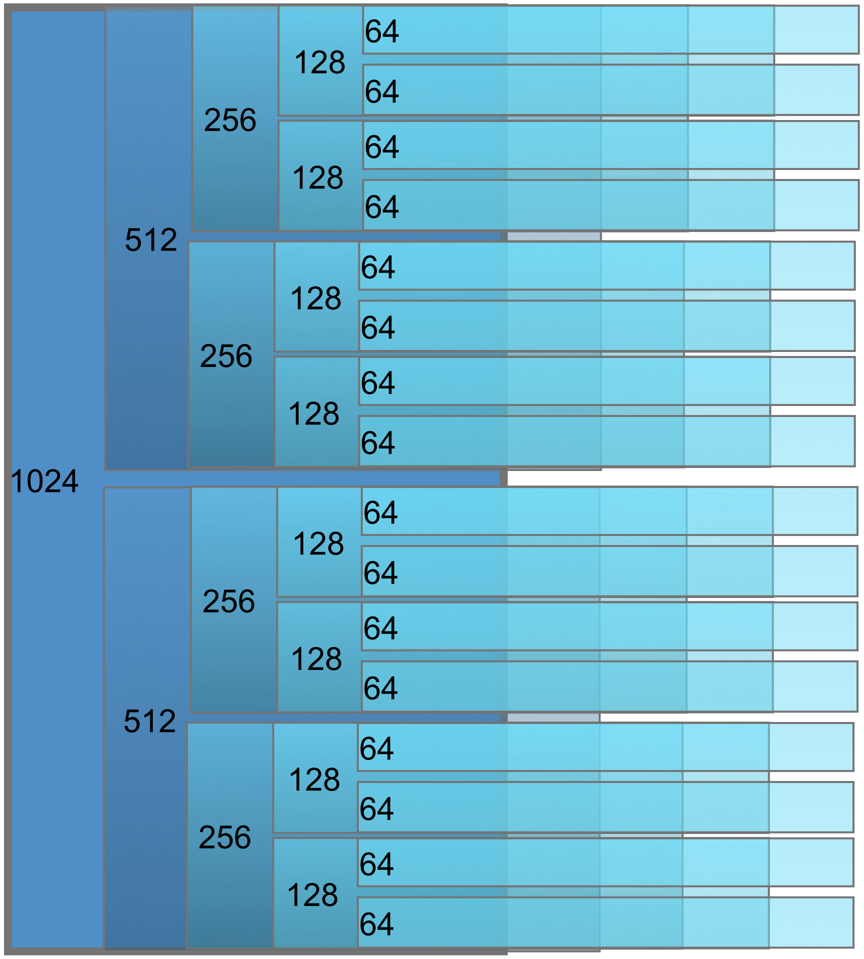
- **Front-end nodes (FN): users login, compile, submit jobs**
  - 2.5 GHz PowerPC 970, Linux OS
- **Service nodes (SN): system management services: create and monitor processes, configure partitions, control jobs, store statistics**
- **I/O nodes (IO): OS services: files, sockets, process management, debugging**
  - *Intrepid*: 1 I/O node per 64 compute nodes
- **Compute nodes (CN): run user applications as batch jobs**
  - CNK OS (no shell)

Blue Gene/P Interconnection Networks



- **3 Dimensional Torus**
  - Interconnects all compute nodes
  - Communications backbone for point-to-point
  - 3.4 Gb/s on all 12 node links (5.1 GB/s per node)
  - 0.5 µs latency between nearest neighbors, 5 µs to the farthest
  - MPI: 3 µs latency for one hop, 10 µs to the farthest
- **Collective Network**
  - Interconnects all compute nodes and I/O nodes
  - One-to-all broadcast functionality
  - Reduction operations for integers and doubles
  - 6.8 Gb/s of bandwidth per link per direction
  - Latency of one way tree traversal 1.3 µs, MPI 5 µs
- **Low Latency Global Barrier and Interrupt**
  - Latency of one way to reach 72K nodes 0.65 µs, MPI 1.6 µs
- **10 Gb/s functional Ethernet**
  - Disk I/O
- **1Gb private control (JTAG)**
  - Service node/system management

Blue Gene/P Partitions



Partitions on 1 rack of *Intrepid*

- ***Intrepid*** **compute nodes grouped into partitions ranging from 64 to 40,960**
  - One I/O node for each 64 compute nodes
- **Job gets entire partition to itself**
- **Minimum partition size is 64 nodes**
- **Each partition is its own torus/mesh**
  - Electrically isolated
  - Rebooted between jobs
- **Partitions <512 nodes form mesh network**
- **Partitions >=512 nodes form torus network**

Programming Environment

- **Languages:**
  - Fortran, C, C++, Python
  - IBM XL and GNU compilers
- **MPI:**
  - Based on MPICH2 1.0.x base code:
    - MPI-IO supported
    - One-sided communication supported
    - No process management (MPI_Spawn(), MPI_Connect(), ….)
  - Uses the 3 different BG/P networks for different MPI functions
- **Threads:**
  - OpenMP 2.5
  - NPTL Pthreads
- **Linux development environment:**
  - Compute Node Kernel provides look and feel of a Linux environment
    - POSIX routines (with some restrictions: no fork() or system())
    - BG/P adds pthread support, additional socket support
  - Statically and dynamically linked libraries

Runtime Environment

- **Three modes for processes per node (one MPI rank per process)**
  - SMP
    - 1 processes accessing all node memory (2 GB)
    - up to 12 threads
  - Dual
    - 2 processes accessing half node memory each
    - up to 6 threads each
  - VN
    - 4 processes accessing one quarter of node memory each
    - up to 3 threads each
- **SPMD model:**
  - Normally, compute nodes all run same executable
  - Alternatives: HTC mode, cobalt-subrun
- **No virtual memory**

Future ALCF System: Blue Gene/Q

- **Evolution of the Blue Gene architecture**
  - 16 cores/node
  - 16 GB of memory/node
  - water cooled
- **Coming in 2012: *Mira***
  - 10 petaFLOPS
  - Over 750K cores
  - 800 TB of memory
  - 70 PB of disk
  - 48 racks
- **BG/P applications should run immediately on the BG/Q**
  - Better performance expected with higher levels of on-node parallelism

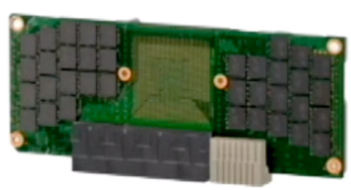
Future ALCF System: Blue Gene/Q

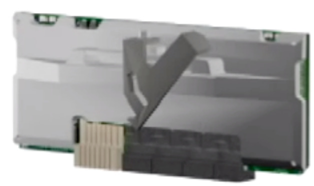
- **Evolution of the Blue Gene architecture**
  - 16 cores/node
  - 16 GB of memory/node
  - water cooled
- **Coming in 2012: *Mira***
  - 10 petaFLOPS
  - Over 750K cores
  - 800 TB of memory
  - 70 PB of disk
  - 48 racks
- **BG/P applications should run immediately on the BG/Q**
  - Better performance expected with higher levels of on-node parallelism

Threads

Blue Gene/Q Packaging *(cont'd)*

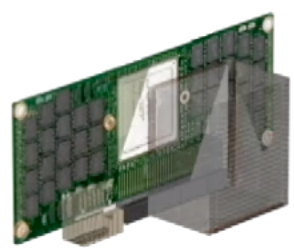
Compute Nodes





Blue Gene/Q Packaging *(cont'd)*

I/O Nodes



Blue Gene/Q Architecture Differences

- **Node-level parallelism**
  - 16 cores
  - 4 hardware threads pre core*
- **Vector/SIMD**
  - Quad FPU*
    - 4-wide double precision FPU SIMD
    - 2-wide complex SIMD
- **Interconnect: 5D torus****
  - Integrates point-to-point, collectives, barriers
  - Supports more flexible mappings of processes onto nodes

*Ruud Haring (IBM), Hot Chips Meeting, 7/2011

**Philip Heidelberger (IBM), Hot Interconnects Meeting, 7/2011

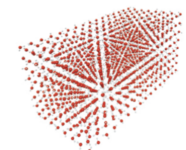
Co-Design of IBM Blue Gene/Q

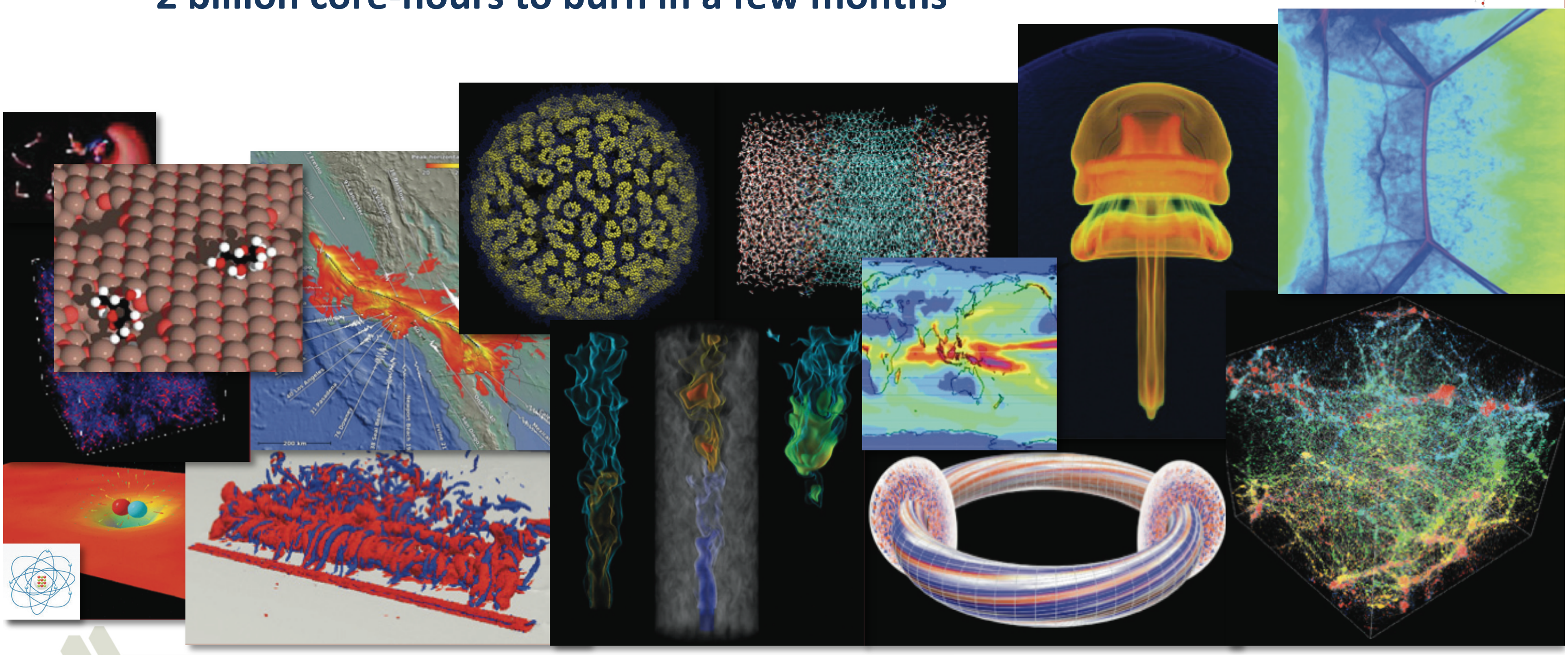
- **Partnership of IBM, LLNL, ANL**
  - Detailed discussions of requirements and hardware/software functionality
    - Quarterly Executive Review meetings
- **Applications and kernels specified in contracts with IBM**
  - Expectations of
    - Functionality
    - Correctness
    - Performance
  - Applications of key importance to labs

First in *Mira* Queue: Early Science Program

- **16 projects**
  - Large target allocation
  - Postdoc
- **Proposed runs between *Mira* acceptance and start of production**
- **2 billion core-hours to burn in a few months**

http://esp.alcf.anl.gov

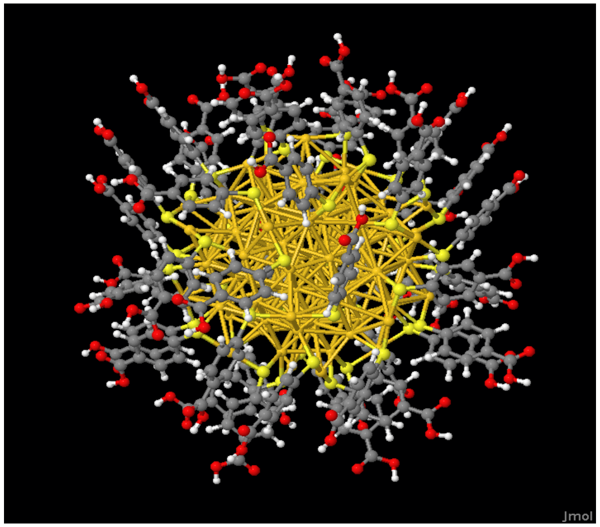




Materials Design and Discovery: Catalysis and Energy Storage

Larry Curtiss, Argonne National Laboratory

- **Electronic Structure Codes: QMCPACK, CPMD**
  - Quantum Monte Carlo (QMC)
  - Density functional theory (DFT)
- **Address catalysis and electric energy storage in 4 areas**
  - Biomass conversion: structure of nanobowls on metal oxide surfaces
  - Electrical energy interfaces
  - Lithium-air batteries
  - Catalysis with transition metal nanoparticles
    - Simulated nanoparticles of up to 1415 atoms using 40% of *Intrepid*



Materials Design and Discovery: Catalysis and Energy Storage *(cont'd)*

- **Quantum Monte Carlo for electronic structure**
  - Operations depend on type of wave function: LCAO, real-space, PWs.
    - Spline interpolation
    - Small DGEMM and DGEMV
- **Current performance**
  - Mixture of compute and bandwidth-limited kernels
    - 5-10% of per core peak performance on IBM Blue Gene/P
    - 20-30% of per core peak performance on x86
  - Heavily rely on C++ compiler optimizations
  - OpenMP 2.5 compliance
- **Paths forward**
  - Reformulate loops to use BLAS2+3 (in progress)
  - Hand tune the SIMD kernels
  - Add nested parallelism to MCWalker evaluation
    - Requires OpenMP 3.0
  - IBM Zurich is optimizing CPMD for Blue Gene/Q

Accurate Simulation of Chemistry in Energy Production & Storage

Robert Harrison, Oak Ridge National Laboratory

- **Codes: MADNESS & MPQC**
- **Catalysis (chemical processes on metal-oxide surfaces)**
  - MADNESS: Model 500-2000 atom lithium oxide clusters
  - MPQC: 50-200 atom models of organic and surface catalysis
    - Run without an eigensolver
- **Heavy element chemistry for fuel reprocessing**
  - Molecular interfacial partitioning
  - Ligand design
  - *Ab initio* dynamics to include finite temperature and entropy

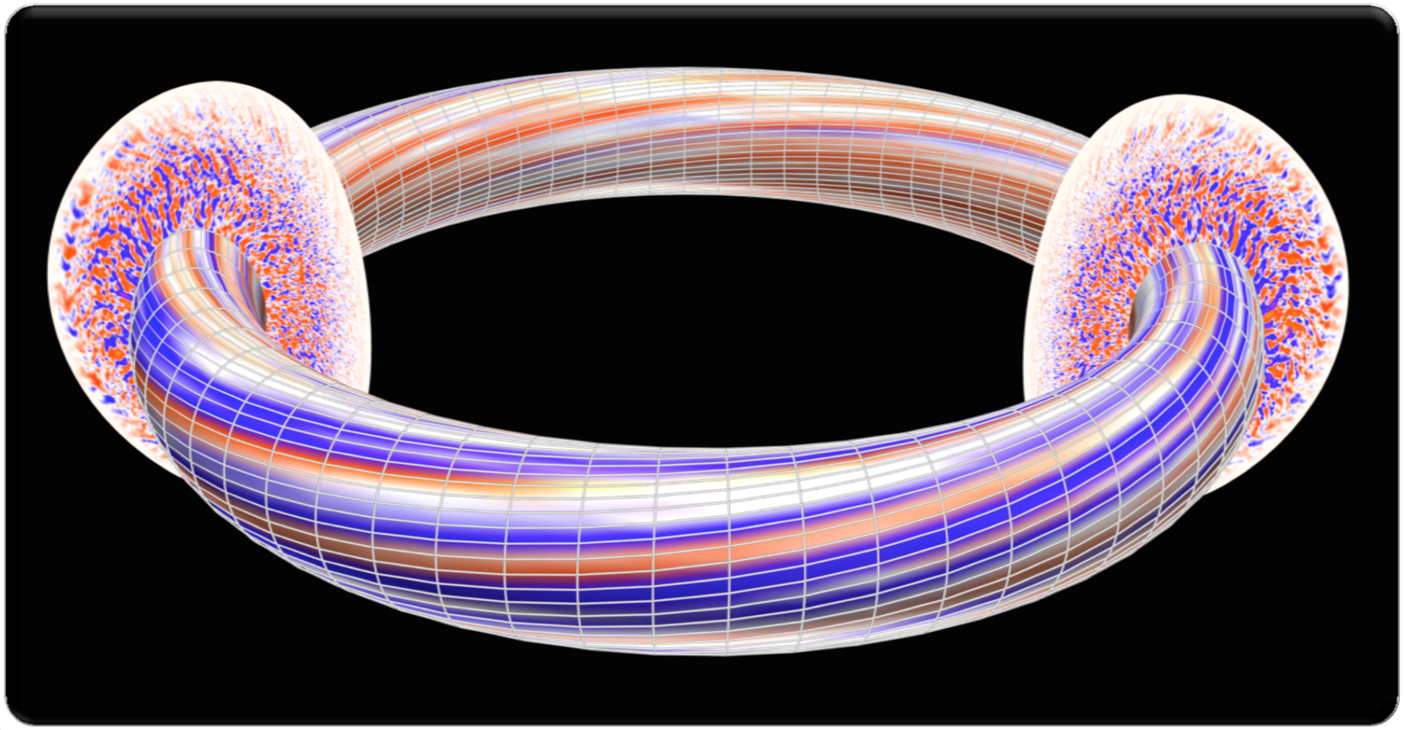
Accurate Simulation of Chemistry in Energy Production & Storage *(cont'd)*

- **Replaced LAPACK with Eigen**
  - fits with C++ OO/template design of MADNESS)
- **Tuning assembly implementation of key kernel (mtxm)**
- **Thread Building Blocks (TBB) port for BGP, POWER7 and BGQ**
- **Pthread + OpenMP interoperability and affinity optimizations underway**
- **Exploring native active-message implementation instead of MPI+polling**

Global Simulation of Plasma Microturbulence at the Petascale & Beyond

William Tang, Princeton Plasma Physics Laboratory

- **Codes: GTC, GTS**
- **Particle-in-cell simulation of plasma**
  - Study energy loss through turbulence
  - Trying to validate key assumption about scaling in ITER



Global Simulation of Plasma Microturbulence at the Petascale & Beyond *(cont'd)*

- **Parallelism: MPI plus loop-level OpenMP**
  - Best Blue Gene/P performance: 1 MPI rank per node with 4 OpenMP threads
  - Best Blue Gene/Q performance: 1 MPI rank per node with 64 OpenMP threads
    - BG/Q has 16 cores/node, 4 hardware threads per core*
    - Running on early-access BG/Q hardware (128 nodes)
  - Mapping ranks to nodes optimizing for 5D network topology

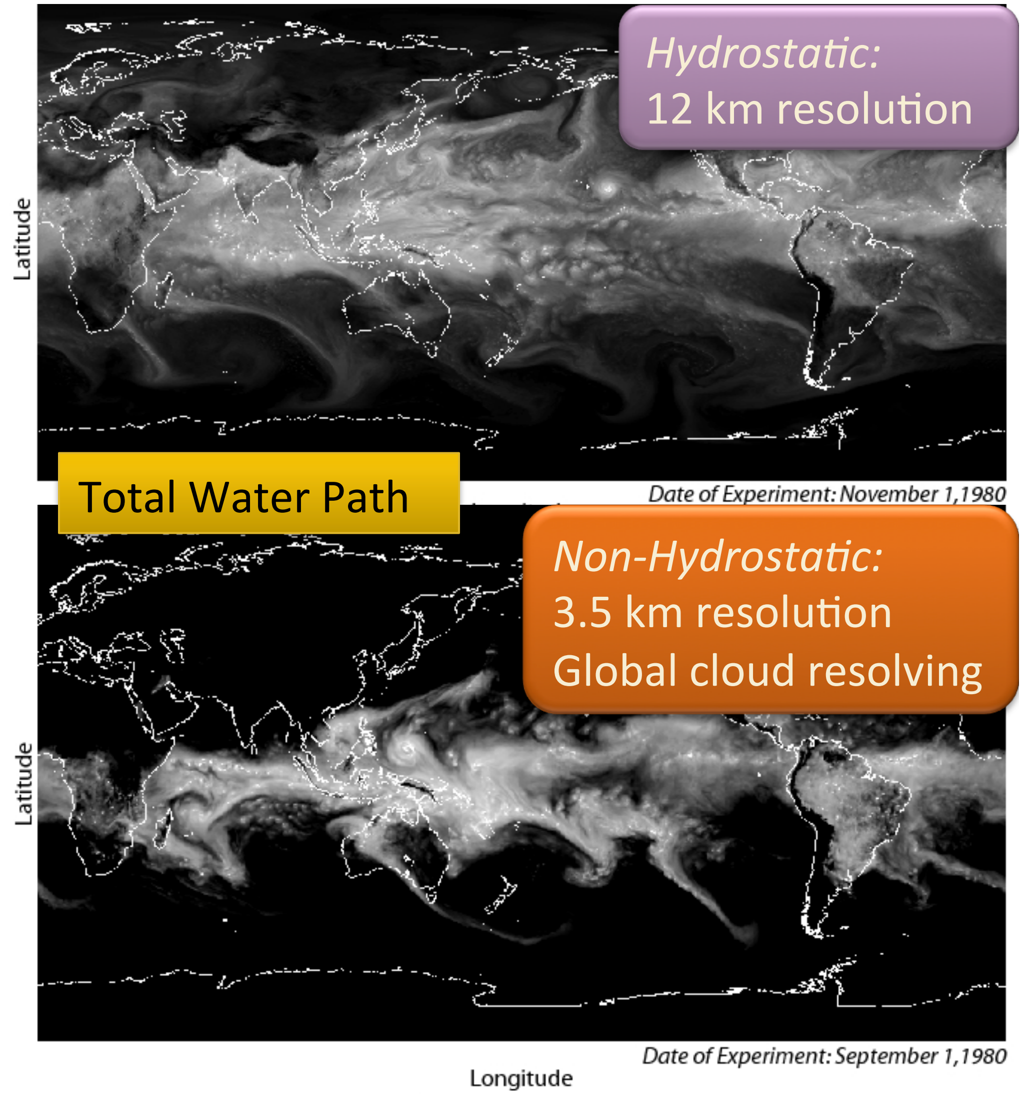
Long-duration simulation of ITER plasmas
- $O(10^{10})$ particles
- $O(10^{8})$ grid cells

*Ruud Haring (IBM), Hot Chips Meeting, 7/2011

Climate-Weather Modeling Studies Using a Prototype Global Cloud-System Resolving Model

V. Balaji, Geophysical Fluid Dynamics Laboratory

- **HIRAM global atmospheric code**
  - Cubed sphere grid
  - *Mira* enables 1st look at effect of clouds on tropical storm statistics
    - Expected to be substantial
  - *Mira* able to run fully-coupled atmosphere and ocean models with resolution resolving clouds



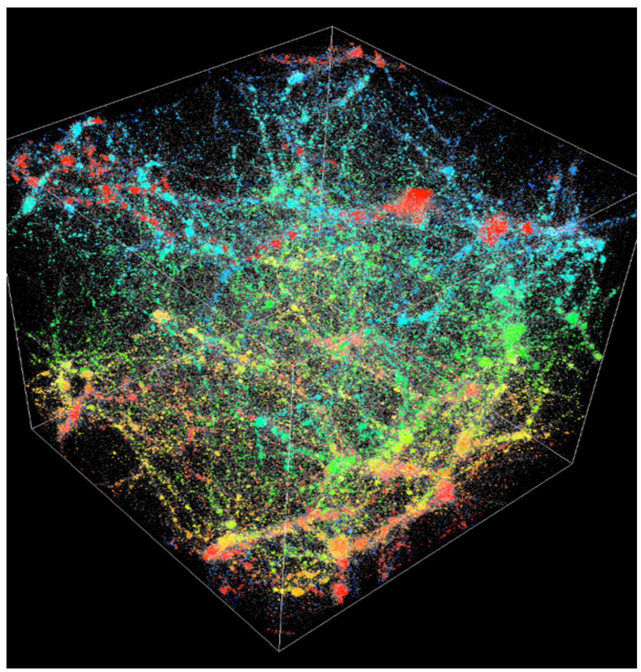
Climate-Weather Modeling Studies Using a Prototype Global Cloud-System Resolving Model *(cont'd)*

- **Flexible Modeling System (FMS) infrastructure**
  - Supports MPI and OpenMP
- **Coarse-grained threads on high-level tasks**
  - Atmospheric physics packages
  - Atmospheric dynamics
  - Land model
- **Loop-level threads used minimally—not efficient**
- **Benchmarking on ANL early access Q machine and other IBM hardware**

Cosmic Structure Probes of the Dark Universe

Salman Habib, Argonne National Laboratory

- **Code: Hardware/Hybrid Accelerated Cosmology Code (HACC)**
- **Gravitational evolution of large-scale structure of the universe**
  - Characterize dark energy & dark matter by predicting observational signatures for a variety of new/existing experimental cosmological probes
    - 1st simulations resolving galaxy-scale mass concentration at scale of state-of-the-art sky surveys
  - Study primordial fluctuations by predicting the effects on cosmic structures today



Cosmic Structure Probes of the Dark Universe *(cont'd)*

- **No change needed for grid layer of code (long-range forces)**
  - Performance of MPI with parallel FFT should be good on Mira
- **Node-level modifications needed (short-range forces)**
  - Rewrite Cell-based code to exploit threads on BG/Q
    - Particle-particle
    - Tree algorithm
- **Hydrodynamics capability**
  - Investigate particle-based methods (hydro-PIC as alternative to SPH)

$O(10^{11}-10^{12})$ grid cells
$O(10^{11}-10^{12})$ particles

Tools and Libraries Project

Kalyan Kumaran, Argonne National Laboratory + 32 co-PIs

- **Performance Tools**
  - PAPI
  - HPCToolkit
  - TAU
  - Scalasca
  - Open|Speedshop
  - FPMPI2
- **Debuggers**
  - DDT (Allinea)
  - TotalView (Rogue Wave)
- **Libraries**
  - FFTW, BLAS
  - PETSc
  - Parallel I/O
    - pNetCDF
    - HDF5
  - Chombo (AMR)
- **Programming Model Implementations**
  - Charm++, AMPI
  - GA Toolkit
  - CoArray Fortran
  - UPC
  - GASnet
  - MPI
- **Visualization Tools**
  - VisIt
  - ParaView

Managed by ALCF, in parallel with ESP projects

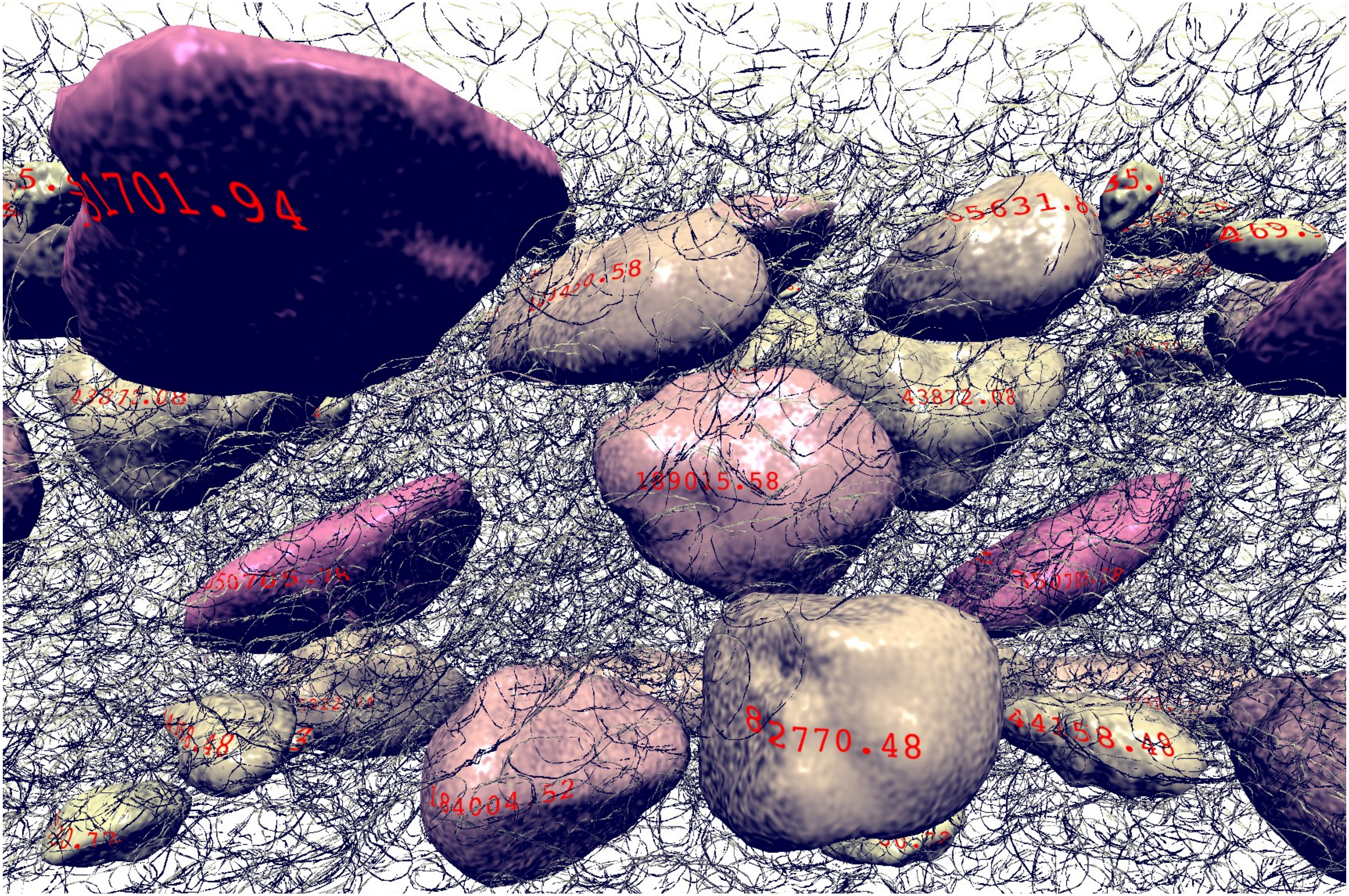
Tools and Libraries Project *(cont'd)*

- **Performance Tools**
  - PAPI
  - HPCToolkit
  - TAU
  - Scalasca
  - Open|Speedshop
  - FPMPI2
- **Debuggers**
  - DDT (Allinea)
  - TotalView (Rogue Wave)
- **Libraries**
  - FFTW, BLAS
  - PETSc
  - Parallel I/O
    - pNetCDF
    - HDF5
  - Chombo (AMR)
- **Programming Model Implementations**
  - Charm++, AMPI
  - GASnet
  - MPI
- **Visualization Tools**
  - VisIt
  - ParaView

HPCToolkit:
Running on Early Access System hardware.
Collecting data for MPI codes.

Tools and Libraries Project *(cont'd)*

- **Global Array (GA) Toolkit**
  - Worked with IBM to provide optimal support for one-sided programming models in PAMI
  - Developing new one-sided communication runtime called OSPRI (One-Sided PRImitives) as replacement for ARMCI on state-of-the-art interconnects (BGQ, PERCS, Gemini)
    - OSPRI aligned with Argonne-led MPI-3 and Unistack efforts, supports a richer set of consistency semantics oriented at application needs and optimal hardware support
    - OSPRI follows prescription for "MPI on a Million Processors," that is, eliminating O(N) algorithms and data structures
  - Reimplementing Global Arrays for hybrid programming models (thread-safety without global lock, internal multithreading, NUMA optimizations)
  - New Global Arrays will support ScaLAPACK as well as 21st-century math libraries (Elemental, PLASMA, MAGMA)


43872.08
189015.58
82770.48